IMPROVING THE CLUSTERING TENDENCY VALUE FOR SYNTHETIC DATA SETS IN CLUSTERING ANALYSIS

POLE ANJAIAH
Institute of Aeronautical Engineering
Email Id: anjaiah.pole@gmail.com

ABSTRACT

The problem of the access tendency plays important role in the clustering analysis. The exact numbers of clusters are detected, but the existing procedure may not work on tight clustered data. Hence, the spectral procedure is developed for achieving quality of results. In large data sets, it requires high computational time, since it depends on Eigen decomposition procedure. In this project, we purpose as new approach (improved version of new clustering tendency approach) and this is implemented on MATLAB 7.0.4 version in windows. The proposed method is derived from the concept shortest paths. The proposed method is tested on various synthetic clustered data sets and concludes that clustering tendency value is improved by proposed methods.

1.INTRODUCTION

The amount of data continues to grow at an enormous rate even though the data stores are already vast. The primary challenge is how to make the data base a competitive business advantage by converting seemingly meaningless data into useful information. How this challenge is met is critical because companies are increasingly relying on effective analysis of the information simply to remain competitive. A mixture of new techniques and technology is emerging to help sort through the data and find useful competitive data.

By knowledge discovery in databases, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases and be investigated from different angles, and large databases thereby serve as rich and reliable sources for knowledge generation and verification. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database systems and machine learning. Companies in many industries also take knowledge discovering as an important area with an opportunity of major revenue. The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications.

From a data warehouse perspective, data mining can be viewed as an advanced stage of on-line analytical processing (OLAP). However, data mining goes far beyond the narrow scope of summarization-style analytical processing of data warehouse systems by incorporating more advanced techniques for data understanding. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases [1].

For example,

1. Learning the application domain
2. Creating a target dataset

3. Data cleaning and preprocessing
4. Data reduction and projection
5. Choosing the function of data mining
6. Choosing the data mining algorithm(s)
7. Data mining
8. Interpretation
9. Using the discovered knowledge

As the KDD process shows, data mining is the central of knowledge discovering, it requires complicated data preparation works. Data cleaning and preprocessing: includes basic operations, such as removing noise or outliers, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding DBMS issues, such as data types, schema, and mapping of missing and unknown values. Useful data are chosen from the formatted data to increase the effectiveness and focus on the task.

2. LITERATURE SURVEY

Mining [1] is one of the fastest growing fields in the computer industry. One of the greatest strengths of data mining is reflected in its wide range of methodologies and techniques that can be applied to a host of problem sets. We can define Data mining as follows.

"Data mining is the process of extracting knowledge from large amounts of data that is usually stored in databases, data warehouses and any other information repositories".

Here the term knowledge means interesting patterns in our data from databases, data warehouses and any other information repositories. A pattern is said to be interesting, if it is easily understood by humans

(i) potentially useful,
(ii) novel ,i.e., Original and of a kind not seen before

Hence we can define Data mining as the "Searching for the interesting patterns in our data".

But what motivated us to move towards data mining?

There is one fundamental reason to have data mining. Now days most of the databases are based on the relations i.e., the data bases are going to store the data in the form of tables. These tables are small enough to analyze, then we can make use of them. But if you consider the real time situation, for example, in super markets or banking systems, there is enormous growth in the data due to their daily transactions. Hence it is tedious task for a human being to analyze such huge amounts of data. We can overcome this problem by applying various data mining techniques.

When we talk about data mining, we may have to go through various important techniques in data mining. They are

(i) Classification
(ii) Clustering
(iii) Association rule mining

The following figure (Fig: 2.1) shows the diagrammatical view of various techniques in data mining.

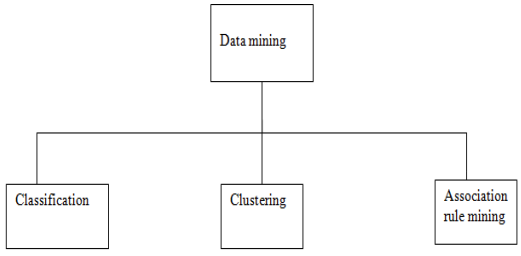


Fig: 2.1 important techniques in data mining.

Data mining is an iterative process, within which progress is defined by discovery, through either automatic or manual methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining [1] is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.

In practice, the two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans.

Therefore, it is possible to put data-mining activities into one of two categories:

- Predictive data mining, which produces the model of the system described by the given data set
- Descriptive data mining, which produces new, nontrivial information based on the available data set.

On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation, or other similar tasks. On the other, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. The relative importance of prediction and description for particular data-mining applications can vary considerably. Data mining has its origins in various disciplines, of which the two most important are statistics and machine learning.

Data mining represents one of the major applications for data warehousing, since the sole function of a data warehouse is to provide information to end users for decision support. Unlike other query tools and application systems, the data mining process provides an end-user with the capacity to extract hidden, nontrivial information. Such information, although more difficult to extract, can provide bigger business and scientific advantages and yield higher returns on "data warehousing and data mining" investments.

3. PARTITIONING METHODS

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

3.1 SPECTRAL VAT

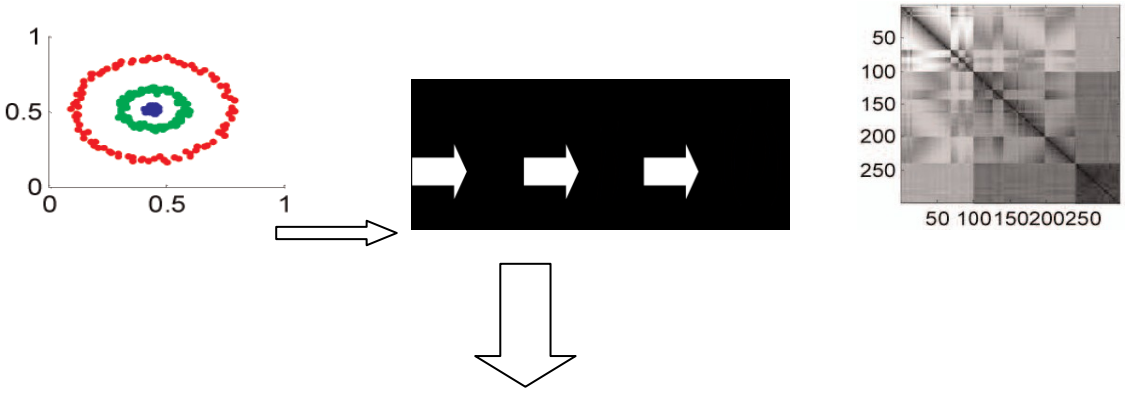
Our work is built upon the VAT [11] algorithm. Two important points about VAT is noted here:

- Only a pair wise dissimilarity matrix D is required as the input. When vectorial forms of objects are available, it is easy to convert them into D using some form of dissimilarity measures. Even when vectorial data are not explicitly available, it is still feasible to use certain flexible metrics to compute a pair wise dissimilarity matrix, e.g., using Dynamic Time Warping (DTW) to match sequences of different lengths.
- Although the VAT image suggests both the number of and approximate members of object clusters, matrix reordering produces neither a partition nor a hierarchy of clusters. It merely reorders the data to reveal its hidden structure, which can be viewed as illustrative data visualization for estimating the number of clusters prior to clustering.

3.2 ALGORITHM STEPS IN SPECTRAL VAT

1. Compute the local scaling parameter oi for object Oi
2. Construct the weighting matrix W
3. Construct the normalized Laplacian matrix L
4. Choose the k largest eigenvectors of L' to form the matrix V
5. Normalize the rows of V with unit Euclidean norm to generate V'
6. Construct a new pair wise dissimilarity matrix D'
7. Apply the VAT algorithm to D'

The spectral decomposition of the Laplacian matrix provides useful information about the properties of the graph. It has been shown experimentally that natural groups in the original data space may not correspond to convex regions, but once they are mapped to a spectral space spanned by the eigenvectors of the Laplacian matrix, they are more likely to be transformed into tight clusters. Based on this observation, we wish to embed D in a k-dimensional spectral space, where k is the number of eigenvectors used, such that each original data point is implicitly replaced with a new vector instance in this new space. After a comprehensive study of recent spectral methods, we adopt a combination of adjacency graph, weighting function, and graph Laplacian for obtaining a better graph embedding.



Spectral Mapping

$$D' \Longrightarrow \tilde{D}' \Longrightarrow \mathrm{I}(\tilde{D}')$$

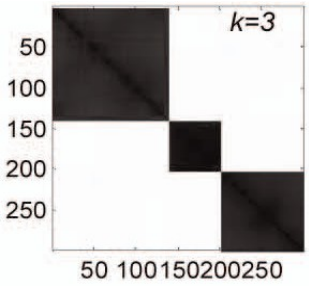


Fig 5.1: Spectral Mapping

Spectral Mapping

$$D \Rightarrow W \Rightarrow L' \Rightarrow V \Rightarrow V' \Rightarrow D'$$

SPECTRAL VAT

O (Kn^2) **O (n^3)**

$$D \Rightarrow W \Rightarrow L' \Rightarrow V \Rightarrow V' \Rightarrow D'$$

$D \Rightarrow W$ (n × n affinity matriz)

$w_{ij} = \exp(-d_{ij}d_{ji}/(\sigma_i\sigma_j))$ for $i \neq j$, and $w_{ii} = 0$

$\sigma_i = d(o_i, o_K) = d_{iK}$ where o_K is the K-th nearest neighbor of o_i

$W \Rightarrow L'$ (normalized Laplacian matrix)

$\boldsymbol{L' = M^{-1/2} W M^{-1/2}}$

M is diagonal, $m_{ii} = \sum_{j=1}^{n} w_{ij}$

$L' \Rightarrow V$ Choose the k largest eigenvectors of L' to form the matrix V

$\boldsymbol{V = [v_1, \cdots, v_k] \in \mathcal{R}^{n \times k}}$

$V \Rightarrow V'$ a new instance (corresponding to o_i)

Normalize the rows of $\boldsymbol{V}$ with unit Euclidean norm

$V' \Rightarrow D'$ $\boldsymbol{d'_{ij} = d(u_i, u_j)}$

Spec VAT O (Kn^2) O (n^3)

E-Spec VAT $D \Rightarrow W \Rightarrow L' \Rightarrow V \Rightarrow V' \Rightarrow D'$

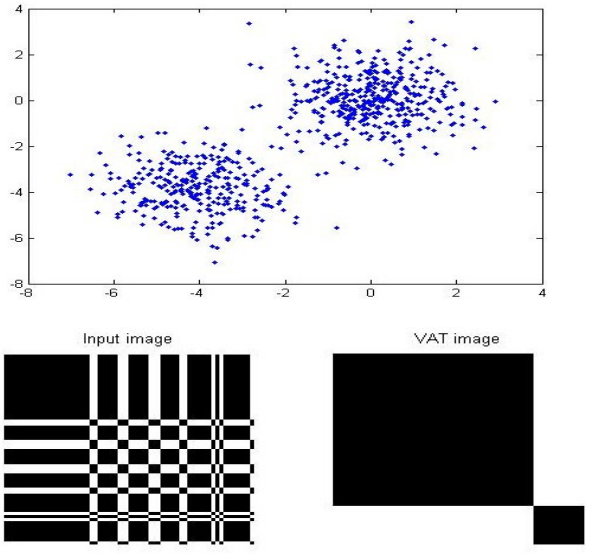
$$D^{n\times n} \Rightarrow (D_S^{m\times m} \; D_B^{m\times(n-m)})$$
$$\Rightarrow \begin{pmatrix} S^{m\times m} & B^{m\times(n-m)} \\ B^T & C \end{pmatrix} (\equiv W)$$

Sampling m (<<n) rows from *D*

$$\tilde{U}_F = \begin{pmatrix} U_S \\ B^T \end{pmatrix}$$

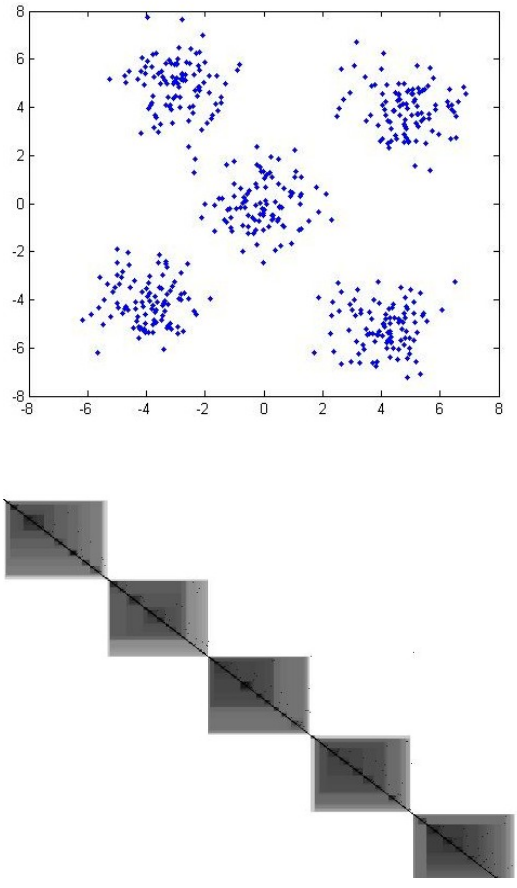
$$\bar{U}_f = \begin{pmatrix} U_S \\ B^T U_S \Sigma_S^+ \end{pmatrix}$$

Two Clustered data and VAT image for two clustered data



We are interested in further exploring the use of image based approaches to assess cluster tendency and extract information about the geometric structure of possible clusters. Questions of interest include finding alternative, superior ordering methods. The method here is closely related to clustering. Cutting the longest connecting edges in produces exactly the single linkage clusters for c = 2 in this data.

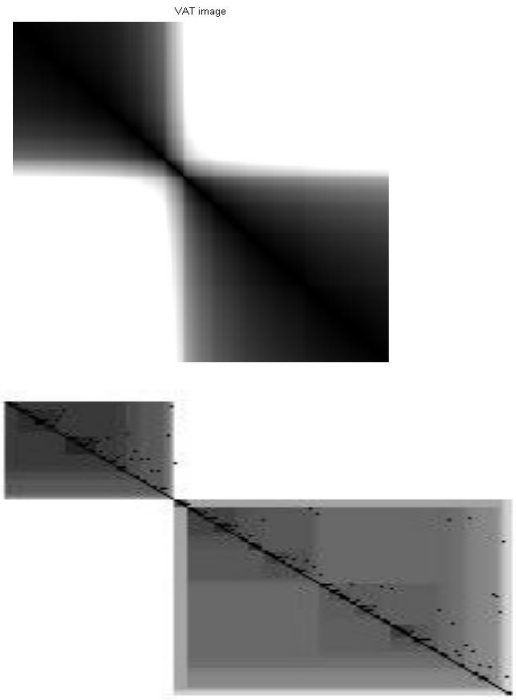
Five Clustered data and VAT image for five clustered data



Five clusters in the data set are visually apparent but there is a high level of mixing

between outliers from components in the mixture. Computing squared Euclidean distance between the pairs of vectors yields a matrix D with dissimilarities for the data set. Accessing only the vectors needed to make a particular distance computation and releasing the memory used by the vectors, to avoid the exhausting memory the processing was broken up calling the extension routine multiple times. We set the number of clusters C=5 for the data set, since running this data set took quite long time the appearance of error rates are small, it is not an easy clustering problem in terms of how well separated the clusters actually are.

Spectral Image for Iris Data and Spectral result for Iris Data+-



We select the best spectral image as the one with the maximum goodness and determine the number of clusters the goodness values and the data depicting the effectiveness of such a measure in determining a good spectral Image. The Computational complexity of this image is mainly depends on computation of Histograms and Optimal Thresholds. Our Visual methods give intuitive Observations on the number of Clusters.

CONCLUSION

We presented an enhanced visual approach towards automatically determining the number of clusters and partitioning data in either object or pair wise relational form, to better reveal the hidden cluster structure, especially for complex-shaped data sets, the VAT algorithm has been improved by using spectral analysis of the proximity matrix of the data. Based on Spectral VAT, a goodness measure of Spectral VAT images has been proposed for automatically determining the number of clusters, derived a visual clustering algorithm based on Spectral VAT images and its unique blocked structured property, and also proposed an extended strategy to scale the Spectral VAT algorithm to larger data sets. A series of primary and comparative experiments on synthetic and real-world data sets have demonstrated that the algorithms perform well in terms of both visual cluster tendency assessment and data partitioning.

The VAT is enhanced by using spectral analysis. Based on Spectral VAT, the cluster structure can be estimated by visual inspection. Number of clusters can be automatically estimated.

BIBLIOGRAPHY

[1] Data Mining: Concepts and Techniques (2nd Edition) by " Han and Kamber"

[2] L.Waing, C.Leckie and J.C.Bezdek, "Automatically Determining the Number of Clusters in Unlabeled Datasets", IEEE Transaction on Knowledge and Data engineering, vol. 21, no. 3, 2009.

[3] T.Havens, J.C.Bezdek, J.Keller and M.Popesu, Dunn.s "Cluster Valid index as a Contrast Measure of VAT Images", IEEE, 2008.

[4] L.Waing , Geng , J.Bezdek and C.Leckie , .Enhanced Visual Analysis for Cluster tendency assessment and Data Partitioning. , IEEE Transaction on Knowledge and Data engineering , vol.22, no.10, pp.1401-1414, 2010.

[5] L.Waing, C.Leckie and J.C.Bezdek, .Automatically Determining the Number of Clusters in Unlabeled Datasets., IEEE Transaction on Knowledge and Data engineering, vol. 21, no. 3, 2009.

[6] T.Havens, J.C.Bezdek, J.Keller and M.Popesu, .Dunn.s Cluster Valid index as a Contrast Measure of VAT Images., IEEE, 2008.

[7] Sanghamitra Bandyopadhyay and Sripama saha, .A Point Symmetry based Clustering Technique for Automatic Evolution of clusters., IEEE Transaction on Knowledge and Data engineering, vol. 20, no. 11, 2008.

[8] I.Sledge, J.Huband and J.C.Bezdek, .(Automatic) Cluster Count Extraction from Unlabeled Datasets., Joint Proceedings Fourth Int.l Conference Natural Computation (ICNC) and Fifth Int.l Conference on Fuzzy Systems and Knowledge discovery (FSKD), 2008.

[9] J.C.Bezdek, R.J.Hathway and J.Huband, .Visual Assessment of Fuzzy Clustering Tendency for Rectangular Dissimilarity Matrices., IEEE Transactions on Systems, vol. 15, no. 5, pp. 890-903, 2007.

[10] L.Waing and Y. Zhang , .On fuzzy cluster validity indices., Fuzzy Sets and Systems, vol. 158, no. 19, pp. 2095-2117, 2007.

[11] R.Hathway, J.C.Bezdek and J.Huband, .Scalable Visual Assessment of Cluster Tendency., Pattern Recognition, vol.39, no. 6, pp. 1315-1324, 2006.

[12] J.Huband, J.C.Bezdek and R.Hathway, .BigVAT: Visual Assessment of Cluster Tendency ., Pattern Recognition, pp. 1875-1886, 2005.

[13] Gautam Garai, B.B.Chaudhuri, .A Novel Genetic Algorithm for Automatic Clustering., Pattern Recognition, Science Direct Letters 25, pp. 173.187, 2004.

[14] U.Maulik and S. Bandyopadhyay, .Performance Evaluation of Some Clustering Algorithms and Validity Indices., IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650-1654, 2002.

[15] J.C.Bezdek and N.R.Pal, .Some new indexes of cluster validity., IEEE Transactions on Systems, Man, And Cybernetics, vol. 28, pp. 301.315, 1998

[16] J.G.Milligan and M.Cooper, .An Examination of Procedures for Determining the Number of Clusters in a Data Set., Psychometrika, vol. 50, pp. 159-179, 1985.

[17] N.Otsu, .A Threshold Selection Method from Gray-level Histograms., IEEE Transaction on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66,1979.

[18] R.F. Ling, .A Computer Generated Aid for Cluster Analysis,.Comm. ACM, vol. 16, pp. 355-361, 1973.

[19] P.Sneath, .A Computer Approach to Numerical Taxonomy., J. General Microbiology, vol. 17, pp. 201-226, 1957.